University of Nevada, Reno

Changes in the Coefficients of Zipf's Law for English Corpora of Different Contexts

A thesis submitted in partial fulfillment
of the requirements for the degree of

Bachelor of Arts in Applied Mathematics and the Honors Program

by

Robert A. Arnold

Dr. Deena Schmidt, Thesis Advisor

December, 2015

UNIVERSITY OF NEVADA RENO

THE HONORS PROGRAM

We recommend that the thesis
prepared under our supervision by

ROBERT ALEXANDER ARNOLD

entitled

Changes in the Coefficients of Zipf's Law for English Corpora of Different Contexts

be accepted in partial fulfillment of the
requirements for the degree of

BACHELOR OF ARTS

_____

Deena Schmidt, Ph. D., Thesis Advisor

_____

Committee Member (if applicable)

_____

Honors Program Representative (if other than Director)

_____

Tamara Valentine, Ph. D., Director, **Honors Program**

December, 2015

Abstract

The statistical behavior of languages has been of great interest to linguists since the mid-20th century. The frequency distribution of words, often modeled with a probability mass function called *Zipf's Law* (Zipf, 1936, 1949), is a particular target of research that has undergone increasingly intense scrutiny over the last decade. It turns out that there is an interesting gap in the research of Zipf's Law pointed out by Steven Piantadosi – namely, language is not static and changes from context to context, and there is comparatively little examination of these fluctuations using Zipf's Law (2014). This paper will set out a course for examining how language changes between different domains of time and different types of written and spoken media via comparing the parameters of best fit for the data using Zipf's Law.

Table of Contents

Abstract ..... **i**
Table of Contents ..... **ii**
List of Tables ..... **iii**
List of Figures ..... **iv**
Introduction **1**
Thesis Question 2
Review of the Literature **3**
Using Zipf's Law to Compare Different Languages 3
Using Zipf's Law to Compare Small Corpora 5
Literature Review Conclusions 6
Methodology **6**
Restatement of the Thesis Question as a Formal Hypothesis 6
Data and Software 7
The First Model 9
The Second Model 12
Results **13**
First Model Tables 13
First Model Graphs 18
First Model Analysis 28
Second Model Tables 30
Second Model Graphs 33
Second Model Analysis 38
Conclusions **38**
References **40**

List of Tables

Table 1

Parameter Estimates, Parameter Error Estimates, and Shapiro-Wilk P-Value 14

Table 2

F Test Ratios, Student T Test Differences for Minimum Rank Cut-Off Value 15

Table 3

F Test P-Values, Student T Test P-Values for Minimum Rank Cut-Off Value.....16

Table 4

F Test Ratios, Student T Test Differences for Exponent of Best Fit.....17

Table 5

F Test P-Values, Student T Test P-Values for Exponent of Best Fit 18

Table 6

Exponent Estimates, Exponent Error Estimates, and Shapiro-Wilk P-Value 30

Table 7

F Test Ratios, Student T Test Differences for Exponent of Best Fit 31

Table 8

F Test P-Values, Student T Test P-Values for Exponent of Best Fit 32

List of Figures

Figure 1: Data Spreadsheet Screenshot **9**
Figure 2: 1800s First Model **20**
(a) Fit (b) alpha Histogram (c) alpha QQ Plot (d) xmin Histogram (e) xmin QQ Plot
Figure 3: 1900-49 First Model **21**
(a) Fit (b) alpha Histogram (c) alpha QQ Plot (d) xmin Histogram (e) xmin QQ Plot
Figure 4: 1950-89 First Model **22**
(a) Fit (b) alpha Histogram (c) alpha QQ Plot (d) xmin Histogram (e) xmin QQ Plot
Figure 5: Academic First Model **23**
(a) Fit (b) alpha Histogram (c) alpha QQ Plot (d) xmin Histogram (e) xmin QQ Plot
Figure 6: Fiction First Model **24**
(a) Fit (b) alpha Histogram (c) alpha QQ Plot (d) xmin Histogram (e) xmin QQ Plot
Figure 7: Magazine First Model **25**
(a) Fit (b) alpha Histogram (c) alpha QQ Plot (d) xmin Histogram (e) xmin QQ Plot
Figure 8: Newspaper First Model **26**
(a) Fit (b) alpha Histogram (c) alpha QQ Plot (d) xmin Histogram (e) xmin QQ Plot
Figure 9: Spoken First Model **27**
(a) Fit (b) alpha Histogram (c) alpha QQ Plot (d) xmin Histogram (e) xmin QQ Plot
Figure 10: 1800s Second Model **34**
(a) Fit (b) alpha Histogram (c) alpha QQ Plot
Figure 11: 1900-49 Second Model **34**
(a) Fit (b) alpha Histogram (c) alpha QQ Plot
Figure 12: 1950-89 Second Model **35**
(a) Fit (b) alpha Histogram (c) alpha QQ Plot
Figure 13: Academic Second Model **35**
(a) Fit (b) alpha Histogram (c) alpha QQ Plot
Figure 14: Fiction Second Model **36**
(a) Fit (b) alpha Histogram (c) alpha QQ Plot
Figure 15: Magazine Second Model **36**
(a) Fit (b) alpha Histogram (c) alpha QQ Plot

Figure 16: Newspaper Second Model.....37
(a) Fit (b) alpha Histogram (c) alpha QQ Plot
Figure 17: Spoken Second Model.....37
(a) Fit (b) alpha Histogram (c) alpha QQ Plot

1 Introduction

Given a corpus of natural language text, the frequency of a word is inversely proportional to its rank in the frequency table. This empirical observation of natural languages is referred to as Zipf's Law (Zipf, 1936, 1949). Though he was not the first person to notice that particular behavior of natural languages, the law is named after George Kingsley Zipf, a linguist who investigated the law and promoted its use in multiple disciplines. Formally stated, Zipf's Law is described by the following relation:

$$f(r) \propto \frac{1}{r^{\alpha}}$$

Here, r is the frequency rank of a given word within a text, $f(r)$ is the frequency, and $\alpha \approx 1$ is a parameter which is determined empirically. In other words, when measuring the frequency of words in a corpus, the second most frequent word (with r = 2) will tend to be about half as frequent as the most frequent word (r = 1), and the third most frequent word (r = 3) will be a third as frequent as the most frequent word, etc…

The statistical behavior of languages has been of great interest to linguists since the mid-20th century. There are two broad areas that have received attention in modern investigations of Zipf's Law: Using Zipf's Law to compare different languages (Gelbukh and Sidorov, 2001), and using Zipf's Law on small corpora such as *The Hound of the Baskervilles* and *Alice in Wonderland* (Baayen, 2001) and restricted categories of words (such as "nouns" and "verbs") within the same language (Marcus, Marcinkiewicz, & Santorini, 1993). These two areas of focus naturally have some overlap in the research – for example, a list of very common words (such as "water", "father", "mother", etc…) that is translated into a sample of world languages (such as English, French, German,

Swedish, etc…) is called a Swadesh List (Calude and Pagel, 2011). Such a list counts as several comparable corpora taken from world languages, but also counts as a category of words that is relatively restricted to content that is universally present in every human society. The key observation to take away from these areas of focus – different languages and small corpora – is that these avenues of research neatly miss the very 'non-staticness' of language that Piantadosi (2014) points out, not examining or taking into consideration change in language over an extended time period or change in language between contexts. For example, the goal of a newspaper is to inform a general audience, while the goal of an academic paper is to inform a much more highly educated general audience or specified experts. These different contexts may result in changes in the word frequencies.

This gap in the research needs to be addressed. These hidden variables could distort the results of other experiments. If a researcher is comparing corpora of text from two languages but fails to control for the time periods from which these corpora draw their respective texts, then is that going to significantly change the results?

1.1 Thesis Question

Given the aforementioned context surrounding the research into Zipf's Law, the question being investigated by this thesis project is the following: What would result in applying Zipf's Law to large corpora organized by time period (for example, a corpus of text taken from the 19th century compared with a corpus of text taken from the time between 1900 and 1950) and by media form (for example, a corpus of text taken solely from newspapers compared with a corpus of text taken solely from magazines)?

2 Review of the Literature

The next section in this paper is divided in three sections. The next two sections will give an overview of research comparing world languages and research comparing small corpora (like *Alice in Wonderland* vs. *The Hound of the Baskervilles*) and restricted categories, respectively, by summarizing and explaining the research that has already been done. The final section will explain how this past research led to the thesis question.

2.1 Using Zipf's Law to Compare Different Languages

In this section, the concentration of research on differences among world languages will be reviewed. The questions answered by these studies ask how Zipf's Law changes depending on what language is used. If the grammatical structure of the language is radically different, does that significantly change the expected results from fitting a power curve like Zipf's Law?

Gelbukh and Sidorov showed that there exists a more than 1% change in the power law exponent for Zipf's Law (2001) – this is referring directly to the "alpha" in the Zipf's Law equation at the beginning of the introduction to this paper. The best-fit coefficient was calculated using a linear regression method employing maximum likelihood estimates (Larsen and Marx, 2006) for a corpus of approximately 2.5 million words of English text vs. a corpus of 2 million words of Russian text. Each word was taken from a wide range of genres including children's books and science fiction. This breadth and depth of the corpora show that there is a measurable difference in the parameter when comparing two different languages and that this difference is not incidental – that is, a 1% difference is very large when taking into account the number of

words and different genres of the corpora which Gelbukh and Sidorov were working with.

More recently, researchers have gone on to explore languages outside of the Indo-European family, such as Chinese, Japanese, and Korean (Lu, *et al*, 2013), noting that thus far, the bulk of research into Zipf's Law and its relationship with natural language have focused on Indo-European languages, as opposed to other language families, such as the Sino-Tibetan family (Chinese) or the Altaic family (Korean). Lu Linyuan's study found that characters taken from these languages exhibited behavior that significantly deviated from the expected power-law pattern following Zipf's Law built from the study of words taken from European languages – notably, that Chinese, Korean, and Japanese characters did not follow Zipf's Law. However, Lu's study did not correctly implement Zipf's Law; the debilitating flaw in that paper lies not in its mathematics, but in the assumption of semantic equality between a word taken from a European language and a Chinese character. Many characters in Chinese form compound words with other characters, behaving much more like morphemes than words. For example, 毕业生 (bì yè shēng, English: graduate) is composed of three characters, each having a meaning (in order: "complete", "industry", and "green") which is unrelated to the whole word. In other words, Chinese characters and words cannot be considered to be comparable to each other.

These studies show that there is no shortage of attention being given to the applications of Zipf's Law to corpora taken from world languages. However, as stated in the introduction of this review, this research does not examine the tendency for language to change over time or for the language to change depending on the context.

2.2 Using Zipf's Law to Compare Small Corpora

For this chapter, studies comparing texts written solely in the same language are discussed – that language specifically being English. What is available to the researcher interested in comparing texts using Zipf's Law when the texts are written in the same language? The set of rules describing the usage of the language suggests itself.

Comparisons have been made between the Zipf's Law parameters (the "alpha" from the equation in the introduction) for categories of words within a language – such as parts of speech (Marcus, Marcinkiewicz, & Santorini, 1993). In their paper, they compared grammatical categories such as determiners, nouns, and verbs in the third person present form (in English), and found that these forms followed Zipf's Law closely, though there were interesting variations. For example, plotting rank versus frequency on a logarithmic scale (so that one increment is ten times as frequent as the next increment down) shows that the curve tends to curve concave down. The presence of a pattern can aid in the task of highlighting abnormal text selections. For example, one deviation from the concave-down behavior observed in Marcus et al's paper – in the case of using verbs to construct a Zipf's Law curve of best fit, is that the curve is concave up instead of down, as noted by Piantadosi in his paper (2014).

That language changes according to context misses the heart of the matter because the corpora described in the previous paragraph are very small or highly restricted, and not representative of an entire language. The coefficients of best fit for a single book may significantly differ from a corpus that draws from a large pool of text.

2.3 Literature Review Conclusions

The research reviewed in this paper is recent, but there is a simple explanation for why there is an abundance of experiments being done now, when looking at the historical context. Because there are now huge, organized databases of texts that can be drawn from, and computers capable of quickly sorting through and producing tables of these data, there is a proliferation of research into the differences between world languages and larger corpora – it only recently became possible due to technological advancements.

All of this research is concerned with the features of language in the *present*, with little concentration on the effect of language evolution over time. Piantadosi (2014) comes close to core of the matter when he talks about social forces changing or adding new words (like "email"), but nowhere does his research mention a systematic study that examines word frequencies over the course of a long period of time. Piantadosi's paper represents the literature; therefore, there is a lack of experiments that test the way Zipf's Law changes over time and between contexts in the research.

3 Methodology

Thesis Question: What would result in applying Zipf's Law to large corpora organized by time period (for example, a corpus of text taken from the 19th century compared with a corpus of text taken from the time between 1900 and 1950) and by media form (for example, a corpus of text taken solely from newspapers compared with a corpus of text taken solely from magazines)?

3.1 Restatement of the Thesis Question as a Formal Hypothesis

The thesis question (What would result in applying Zipf's Law to large corpora organized by time period and by media form?) is too general to stand as a hypothesis; the

question needs to be restated in a form that is statistically testable. Hence, the following formulation is proffered: 1) The null hypothesis is that the differences in the parameters of the Zipf's Law fit between the corpora organized by time period will not be more or less pronounced than the differences in the parameters determined for the corpora organized by media form, 2) the first alternative hypothesis is that the differences for the time-period data will be more pronounced as quantified by student t-tests and F-tests, and 3) the second alternative hypothesis is that the differences for the media-form data will be more pronounced. The methods outlined below are aimed at determining which two of the three hypotheses ought to be rejected.

3.2 Data and Software

For this research project I purchased data from Brigham Young University's CORPORA, specifically the word frequency data for 100,000 words from the Corpus of Contemporary American English (COCA) (Davies 2008) and the Corpus of Historical American English (COHA) (Davies 2010). From the COCA website: "[The] COCA includes 440 million words taken from 190,000 texts during 1990-2012, evenly divided (~88 million words each) into spoken, fiction, magazine, newspaper, academic." From the COHA website: "The COHA data includes 385 million words of text in 116,000 different texts from the 1810s-2000s, in fiction, popular magazines, newspapers, and non-fiction (books)." For this paper I did not investigate how the data was collected or question the integrity of the data. Since it is one of the most extensive databases available to date, many graduate students and professors use the corpus, a list of which can be seen on the corpus website at (http://corpus.byu.edu/researchers.asp).

These data are organized into eight sub-corpora which I have grouped into two main categories: time-period data and media-form data. The time-period data consist of words taken from texts from the 1800s, 1900-1949, and 1950-1989, each category without regard for different media forms. The media-form data consist of words taken from newspapers, magazines, fiction literature, academia, and spoken (speech from television or radio), each category specifically taken from between 1990-2012. These data are uploaded into the R statistical software language (R Core Team, 2015), with the eight sub-corpora represented in R as separate "data-frame" objects, each with two columns: the frequency ranks of given words, and the frequencies of those words.

Then, for each corpus, any word which had zero tokens (zero "appearances" in the corpus) was removed from the list. The rationale for not including zero-frequency words is simple: If a word doesn't appear in the corpus, then creating a model which notes its absence is a strange thing to do. For example, "e-mail" is, of course, a word that doesn't appear in the 1800s. When building a model for words taken from texts from the 1800s, why have a data point that says "e-mail" appeared zero times? To be fair, there would then be a need to have such a data point for *every English word* which is not present in the 1800s corpus. For this reason, I decided to discard any word which doesn't appear.

On the next page is a screenshot of the spreadsheet from which the data used in the models are drawn:

ID	w1	L1	c1	CAPS	US/UK	freq	COCA	BNC	SOAP	1950-89	1900-49	1800s	coca_spok	coca_fic	coca_mag	coca_news	coca_acad
1	the	the	at	0.11		2.5E+07	54124.71	59717.97	21403.42	59363.87	63479.96	65266.92	46393.26	53301.68	53775.83	53613.78	63981.74
2	and	and	cc	0.08		1.2E+07	26636.9	25808.3	17677.4	26260.1	28577.4	33417.5	26089.7	25756	26458.2	24577.2	30346.6
3	of	of	ii	0.01		1.2E+07	25782.8	30086.5	10067.3	27505.5	32184.2	37182.9	21502.9	19640.2	25872.2	23814	38261
4	a	a	at	0.06		1E+07	22240.8	20853.5	15632.5	22557.5	21357.8	20346.3	21403.6	22960.8	24395.4	23734.4	18637.5
5	in	in	ii	0.09		8035789	17306.2	18307.5	6702.35	17055.2	17335.6	17775.9	15433.9	13195	17503.2	18490.8	21952.5
6	to	to	to	0.02		7277326	15672.8	15564.9	22985.9	14966.6	14749.6	14833.1	18518.5	14851.2	15252.3	15091.2	14527.9
7	I	i	pp	1.00		4725236	10176.5	7797.46	49055.5	10999.8	10607.5	10672.6	17759.3	18076.8	7119.85	5617.46	2174.98
8	to	to	ii	0.02		4456281	9597.23	10080	7763.66	10231.8	11044.5	11980.1	8584.14	9325.53	9766.66	9376.73	10973.7
9	it	it	pp	0.24		4453425	9591.08	10402.4	18411.5	10366.9	11081.1	10752.7	14631.5	11932	8239.03	7851.97	5148.77
10	is	be	vbz	0.02		4237443	9125.93	9780.17	9837.83	7973.66	8449.09	8909.37	12536.6	5146.32	9019.08	8862.46	9875.33
11	that	that	cs	0.07		3921414	8445.32	7303.43	8529.15	7642.72	8159.12	8173.44	11722	5711.03	7831.18	7256.74	9563.24
12	for	for	ii	0.07		3787585	8157.1	8253.02	7319.81	7560.57	7533.73	7024.17	7459.64	6309.93	8729.55	9218.25	9053.34
13	you	you	pp	0.17		3568520	7685.31	6614.9	51035.2	7355.25	7314.26	6036.75	17140.2	10761.6	6009.35	3199.71	986.81
14	was	be	vbdz	0.01		3384649	7289.32	8725.73	6909.42	10092.2	10602.5	9767.91	7800.68	12102.3	5568.66	6017.86	5058.96
15	he	he	pp	0.30		3323458	7157.53	6333.89	7726.87	11817	11511.2	9773.35	7400.62	14184.8	5192.28	7055.8	2089.12
16	with	with	ii	0.04		3095076	6665.68	6508.3	6098.61	6628.54	7011.68	7935.57	5720.07	6784.76	7358.32	6673.06	6808.19
17	on	on	ii	0.05		2871374	6183.91	6417.23	3916.7	5997.28	5348.41	4625.34	6122.29	6409.43	6326.31	6397.21	5660.04
18	's	's	ge	0.00		2439692	5254.22	4362.83	2752.64	4881.43	3999.74	3361.87	3096.59	5246.68	6039.49	6967.15	4976.35
19	at	at	ii	0.08		2275149	4899.85	5165.15	2880.69	5395.61	5504.03	5205.96	4242.86	5950.3	4958.05	5697.53	3683.32

Figure 1: Data Spreadsheet Screenshot

The leftmost column here is the frequency rank of the word in the second column. For example, the word “of” has a frequency rank of 3 – it is the third most frequent word overall. The orange highlighted column is the corresponding overall raw frequency (the number of occurrences) of the word in the COCA. The three lighter-yellow columns with the respective headings 1950-89, 1900-49, and 1800s are the per-million word frequencies for those time-period categories taken from the COHA. The five green columns that are headed respectively with coca_spok, coca_fic, coca_mag, coca_news, and coca_acad are the per-million word frequencies for those media-form categories taken from the COCA.

3.3 The First Model

Zipf’s Law, explicitly stated (Zipf, 1936, 1949):

$$f(r) \propto \frac{1}{r^{\alpha}}$$

Here, r is the frequency rank of a given word within a text, $f(r)$ is the frequency, and $\alpha \approx 1$ is a parameter which is determined empirically. In other words, when measuring the frequency of words in a corpus, the second most frequent word (with r = 2)

will tend to be about half as frequent as the most frequent word (r = 1), and the third most frequent word (r = 3) will be a third as frequent as the most frequent word, etc…

Zipf's Law is a special case of a power law distribution. This distribution is typically shown graphically on a double logarithmic plot as a straight line. A double logarithmic plot is one in which, instead of graphing the horizontal and vertical coordinates for points directly, the logarithm of the horizontal coordinates and the logarithm of the vertical coordinates are used instead.

The main functions used to produce the model are in the "poweRlaw" package (Gillespie 2015), which I installed into R. I calculated the coefficients of the Zipf power laws fitting the various sub-corpora using the method outlined in *Power-Law Distributions in Empirical Data* (Clauset 2015). There's a slight variation from the original Zipf's Law: Clauset calls for not only fitting a power law curve to the data, but also for determining the minimum rank beyond which the data most characterizes a power curve:

$$f(r) = \frac{C}{r^{\alpha}}, r > x_{\min}$$

$$C = (\alpha - 1)x_{\min}^{\alpha-1}$$

Where α (alpha) is the exponent of the power law, $x_{\min}$ (xmin) is the minimum rank cut-off value, and C is a normalization constant specifically meant to keep the sum of proportions equal to 1. Below is a summarization of the steps in the modelling process:

- First, point estimates of the discrete power law exponent and the least-ranked data point which produces the minimum sum of squared errors will be calculated for each sub-corpus. This will be done in R using the functions "displ" and

"estimate_xmin". The functions for fitting the model work by finding the best fit power law exponent (parameter "alpha" (α)) via a maximum likelihood function and finding the minimum rank cut-off value (parameter "xmin" ($x_{\min}$)) through minimizing the Kolmogorov-Smirnoff statistic (Chakravarti 1967). This numerical process is initialized with the following MLE, outlined by Gillespie (2015) in his R package:

$$\hat{\alpha} \simeq 1 + n\left[\sum_{i=1}^{n} \ln \frac{x_i}{x_{\min} - 0.5}\right]^{-1}$$

- Then the standard deviation of the uncertainty of these estimates of the power law exponent and the cut-off value are measured in a bootstrap procedure via the functions "bootstrap" and "sd". A histogram and a quantile-quantile plot of 1000 bootstrap iterations for each parameter will be produced.
- The relative normality of the bootstrap iteration data is tested using a Shapiro-Wilk test (Larsen and Marx, 2006). The P-values of the Shapiro-Wilk tests show how closely a set of data follows a normal distribution – if the P-value is above 5%, then it is reasonable to assume that the data is normal. It is desirable for the bootstrap iterations to follow a normal distribution because otherwise a comparison between the iterations for two different parameters is rendered nonsensical. The F tests and the student-t tests depend upon the data being compared having a normal distribution.
- The variance ratios between the parameters of each pair of corpora are tested using F tests. The hypothesis of the F tests is that the variances have a ratio of 1. If the P-value of an F-test is smaller than 5%, then that is significant evidence that

the variances of the two parameters differ. By itself, a difference in variances can be a metric for deciding that two parameters are different from each other.

- Finally, these results are used to test the hypothesis that the parameters differ significantly from one another. This is done through the use of a student-t distribution hypothesis test (Lehmann 1993) for each parameter for each pair of corpora, using a 95% tolerance level.

3.4 The Second Model

In the process of producing the first model, several problems arose. Specifically, Clauset's algorithm (2015) had a difficult time deciding the optimal minimum rank cut-off value because the data points slope downwards with a gentle curve before straightening, and the optimal minimum rank cut-off value tended to be very large. In addition, the variance in the bootstrap iterations was too high and the selections were not normally distributed according to the Shapiro-Wilk tests, making a comparison between the calculated parameters very difficult to accomplish. As a result, I decided to modify the method in order to perform the statistical analysis. A full explanation of the rationale for these changes is given in the First Model Analysis section on p. 28.

Firstly, instead of allowing the minimum rank cut-off value to float and be decided by the algorithm, I chose a value for the parameter for each corpus such that only the 3500 most frequent words in each corpus be modeled. In this way, the uncertainty in the minimum rank cut-off value is removed from the calculation and the exponent values are able to be compared with each other. Secondly, in addition to setting the minimum rank cut-off value for each corpus, the bootstrap procedure was run for 5000 iterations

instead of 1000 for this second set of models in order to better ensure that the bootstrap iterations are normally distributed.

Other than that, I repeat the same analysis that was done in the first model here: Shapiro-Wilk tests, F-tests, and student-t tests for the alpha parameters.

4 Results

Thesis Question: What would result in applying Zipf's Law to large corpora organized by time period (for example, a corpus of text taken from the 19th century compared with a corpus of text taken from the time between 1900 and 1950) and by media form (for example, a corpus of text taken solely from newspapers compared with a corpus of text taken solely from magazines)?

Formal Hypothesis: The null hypothesis is that the differences in the parameters of the Zipf's Law fit between the corpora organized by time period will not be more or less pronounced than the differences in the parameters determined for the corpora organized by media form. The first alternative hypothesis is that the differences for the time-period data will be more pronounced as quantified by F-tests and student t-tests, and the second alternative hypothesis is that the differences for the media-form data will be more pronounced. The methods outlined below are aimed at determining which two of the three hypotheses ought to be rejected.

4.1 First Model Tables

The tables of the results of the tests and calculations for the first model are in this section. A short description of the data presented in each table is given after each caption:

Corpus	Minimum Rank Cut-Off Value	Standard Error 95%	Shapiro-Wilk Test P-Value	Exponent of Best Fit	Standard Error 95%	Shapiro-Wilk Test P-Value
1800s	1221	239.6	<2.2e-16	1.883	0.0299	3.29e-12
1900-49	1083	419.3	<2.2e-16	1.860	0.0468	1.80e-10
1950-89	961	359.3	<2.2e-16	1.851	0.0490	<2.2e-16
Academic	2395	712.3	<2.2e-16	1.940	0.0625	<2.2e-16
Fiction	665	211.1	<2.2e-16	1.799	0.0286	1.96e-14
Magazine	1130	337.6	<2.2e-16	1.883	0.0381	4.06e-10
Newspaper	1065	601.6	<2.2e-16	1.841	0.0589	<2.2e-16
Spoken	628	257.9	<2.2e-16	1.728	0.0462	<2.2e-16

Table 1: Parameter Estimates, Parameter Error Estimates, and Shapiro-Wilk P-Value

The first column is the name of the corpus. The second column gives the point estimate of what the minimum rank cut-off value should be using Clauset's algorithm's power-law fit (2015) for the given corpus. The next column is the number that should be added or subtracted in order to get the 95% confidence interval for the point estimate (for example, for the Magazine corpus, the calculated minimum rank cut-off value(the "xmin" parameter) is 1130 plus or minus 337.6). The fourth column is the P-value for the Shapiro-Wilk Test for normality – a low P-value (less than 5%) means the null hypothesis that the parameter estimate bootstrap iterations are normal can be rejected. In other words, a P-value is less than 5% indicates that the estimates do not follow a normal distribution. Columns five through seven tell the same information, except for the exponent of best fit (the "alpha" parameter).

Time-Period Corpus Pair	F Test Variance Ratio	Student T Test Mean Absolute Difference
1800s vs 1900-49	0.3266	26.25
1800s vs 1950-89	0.4447	31.65
1900-49 vs 1950-89	1.3617	5.4
Media-Form Corpus Pair	F Test Variance Ratio	Student T Test Mean Absolute Difference
Academic vs Fiction	11.386	1387.8
Academic vs Magazine	4.4520	963.6
Academic vs Newspaper	1.4019	816
Academic vs Spoken	7.6277	1503
Fiction vs Magazine	0.3910	424.2
Fiction vs Newspaper	0.1231	571.8
Fiction vs Spoken	0.6699	115.2
Magazine vs Newspaper	0.3149	147.6
Magazine vs Spoken	1.7133	539.4
Newspaper vs Spoken	5.4409	687

Table 2: F Test Ratios, Student T Test Differences for Minimum Rank Cut-Off Value

In this table there are results of tests which directly compare the minimum rank cut-off value (the "xmin") of one corpus to that of another corpus. For this thesis, I compared a time-period corpus only with another time-period corpus, and I compared a media-form corpus only with another media-form corpus. In this way, I obtain measures and statistics for how different the time-period corpora are from each other and for how different the media-form corpora are from each other.

There are six sections in the table. The top three sections represent the tests performed between the time-period corpora data, and the bottom three sections represent the tests performed between the media-form corpora data.

The left-most column is the name of the corpora-pair which is being tested. Here, one sees that there are three possible pairs with the time-period data since there are only three time-period corpora (1800s, 1900-49, and 1950-89) and that there are ten such pairs with the media-form corpora since there five total media-form corpora.

The middle column is the ratio between the variances in the bootstrap iteration data for each corpus. For example, in the row with the name "Academic vs. Spoken", the number in the second column is calculated by first calculating the sample variance of the Academic bootstrap iteration data, then by calculating the sample variance of the Spoken bootstrap iteration data, and then by dividing the Academic sample variance by the Spoken sample variance. If the number that results is close to 1, then that means that the variances are nearly equal to each other.

The right-most column is the absolute value of the difference between the estimated means of the two bootstrap iteration data for the two corpora. For example, with the "Academic vs Spoken" data, first the sample mean of the bootstrap iterations for the Academic data and for the Spoken data are calculated and then the distance between them is measured – that's the number in the third column. If that distance is close to zero, then that means that the sample means are close to each other.

Time-Period Corpus Pair	F Test P-Value	Student T Test P-Value
1800s vs 1900-49	<2.2e-16	0.0858
1800s vs 1950-89	<2.2e-16	0.0206
1900-49 vs 1950-89	1.13e-06	0.7571
Media-Form Corpus Pair	**F Test P-Value**	**Student T Test P-Value**
Academic vs Fiction	<2.2e-16	<2.2e-16
Academic vs Magazine	<2.2e-16	<2.2e-16
Academic vs Newspaper	1.01e-07	<2.2e-16
Academic vs Spoken	<2.2e-16	<2.2e-16
Fiction vs Magazine	<2.2e-16	<2.2e-16
Fiction vs Newspaper	<2.2e-16	<2.2e-16
Fiction vs Spoken	2.82e-10	<2.2e-16
Magazine vs Newspaper	<2.2e-16	1.856e-11
Magazine vs Spoken	<2.2e-16	<2.2e-16
Newspaper vs Spoken	<2.2e-16	<2.2e-16

Table 3: F Test P-Values, Student T Test P-Values for Minimum Rank Cut-Off Value

Table 3 corresponds to Table 2, except that instead of the calculated statistic, the P-value of the test is given instead.

When the P-value of the F-test is low (less than 5%) it means that the null hypothesis of the variances of the bootstrap iterations being equal to each other (the hypothesis that the variance ratio from the second column of Table 2 is equal to 1) can be safely rejected. In the second column, one can see that all the P-values are very low. That means that for every pair, the variances are significantly different from each other.

When the P-value of the student-t test is low (less than 5%) it means that the null hypothesis of the means of the bootstrap iterations being equal to each other (the hypothesis that the absolute difference from the third column of Table 2 is equal to zero) can be safely rejected. Except for the 1800s vs 1900-49 pair and the 1900-49 vs 1950-89 pair, all of these P-values are less than 5%, meaning that the means of the bootstrap iterations can be taken to be significantly different from each other.

Time-Period Corpus Pair	F Test Variance Ratio	Student T Test Mean Absolute Difference
1800s vs 1900-49	0.4093	0.0174
1800s vs 1950-89	0.3740	0.0150
1900-49 vs 1950-89	0.9137	0.0024
Media-Form Corpus Pair	**F Test Variance Ratio**	**Student T Test Mean Absolute Difference**
Academic vs Fiction	4.7776	0.1127
Academic vs Magazine	2.6928	0.0351
Academic vs Newspaper	1.1250	0.0578
Academic vs Spoken	1.8297	0.1971
Fiction vs Magazine	0.5636	0.0776
Fiction vs Newspaper	0.2355	0.0549
Fiction vs Spoken	0.3829	0.0844
Magazine vs Newspaper	0.4178	0.0227
Magazine vs Spoken	0.6795	0.1621
Newspaper vs Spoken	1.6263	0.1394

Table 4: F Test Ratios, Student T Test Differences for Exponent of Best Fit

Table 4 is like Table 2, except that all of the calculations are being done for the exponent of best fit ("alpha") instead of the minimum rank cut-off value.

Time-Period Corpus Pair	F Test P-Value	Student T Test P-Value
1800s vs 1900-49	<2.2e-16	<2.2e-16
1800s vs 1950-89	<2.2e-16	2.547e-16
1900-49 vs 1950-89	0.1537	0.2585
Media-Form Corpus Pair	**F Test P-Value**	**Student T Test P-Value**
Academic vs Fiction	<2.2e-16	<2.2e-16
Academic vs Magazine	<2.2e-16	<2.2e-16
Academic vs Newspaper	0.06278	<2.2e-16
Academic vs Spoken	<2.2e-16	<2.2e-16
Fiction vs Magazine	<2.2e-16	<2.2e-16
Fiction vs Newspaper	<2.2e-16	<2.2e-16
Fiction vs Spoken	<2.2e-16	<2.2e-16
Magazine vs Newspaper	<2.2e-16	<2.2e-16
Magazine vs Spoken	1.15e-09	<2.2e-16
Newspaper vs Spoken	2.07e-14	<2.2e-16

Table 5: F Test P-Values, Student T Test P-Values for Exponent of Best Fit

This table is related to Table 4 in the same way that Table 3 is related to Table 2: These are the P-values for the F-tests and the student-t tests. The P-values have the same meaning as in the explanation for Table 3. In a similar fashion, most of the P-values are very low.

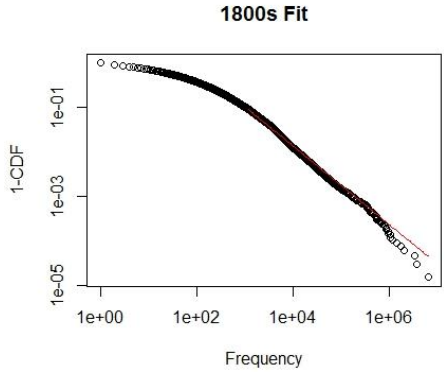
4.2 First Model Graphs

The graphical results of the first model are shown in Figures 2-9. The figures are organized in the following way: Each figure consists of five parts per page representing information about exactly one corpus. For example, Figure 2 focuses on the 1800s corpus.

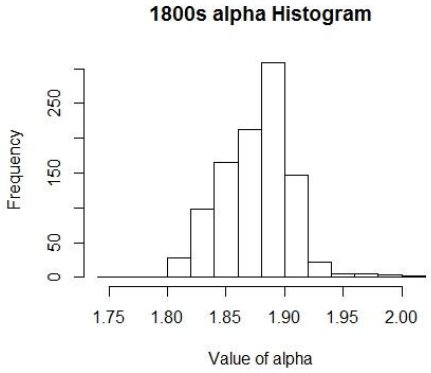
Part (a), the top image, is a plot of the data along with the model fit. The black circles are data points, and the solid black is just a very dense cluster of data points. The red line is the model fit.

Parts (b) and (c) are graphs showing information about the bootstrap iterations of the exponent parameter “alpha”. Part (b) is a histogram showing the bins of the values calculated at each of the 1000 iterations. The most desirable shape of the histogram is a “bell-curve” because that means the iterations are approximately normally distributed, and normality is desired in order to be able to compare the alpha parameter for the corpus with the alpha parameter for the other corpora. Part (c) shows the same data in a different format, called a normal quantile-quantile (QQ) plot, which pairs each data point’s percentage quantile with the quantile of a point taken from a normal distribution. Here, the desirable thing to have is for all the data points to be as close to the straight line as possible, which would, in the same way as a “bell-curve” in a histogram, show normality.

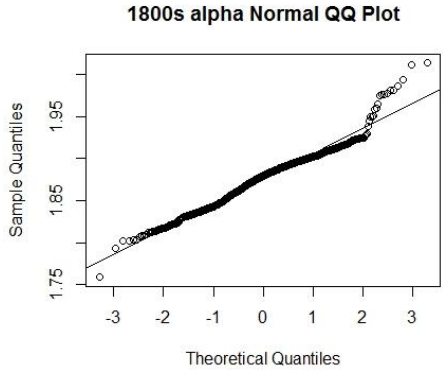
Parts (d) and (e) are for the minimum rank cut-off parameter “xmin”. They show the analogous information as (b) and (c) do, with a histogram and a normal QQ plot of the bootstrap iterations.



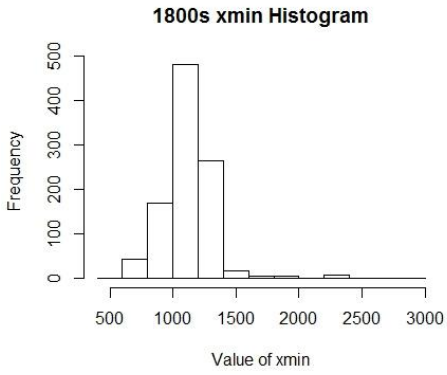
(a)



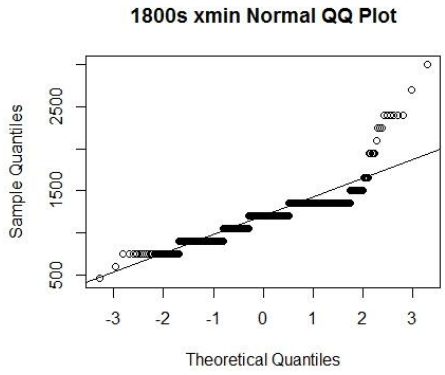
(b)



(c)

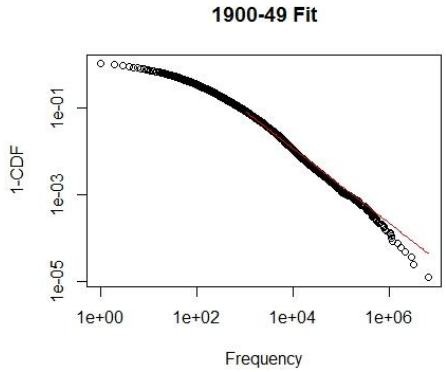


(d)

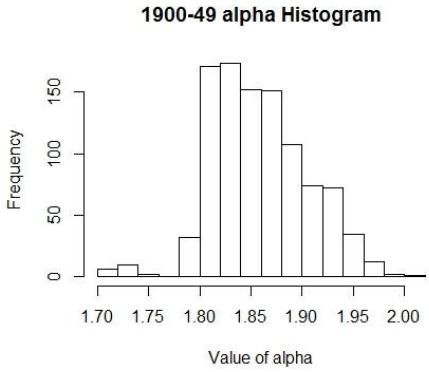


(e)

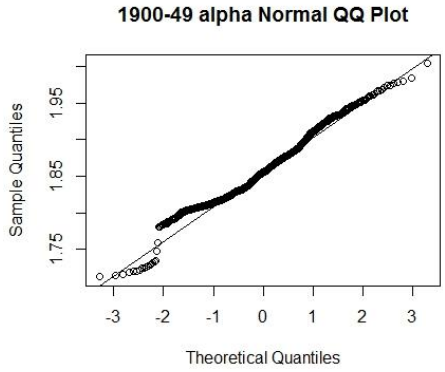
Figure 2: 1800s First Model. (a) Fit (b) alpha Histogram (c) alpha QQ Plot (d) xmin Histogram (e) xmin QQ Plot



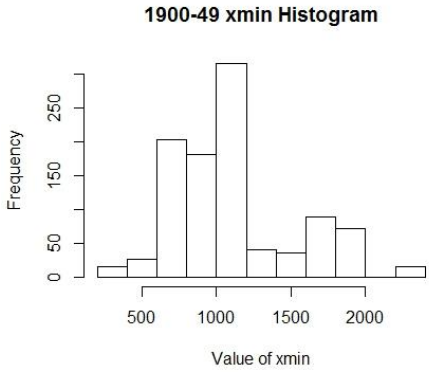
(a)



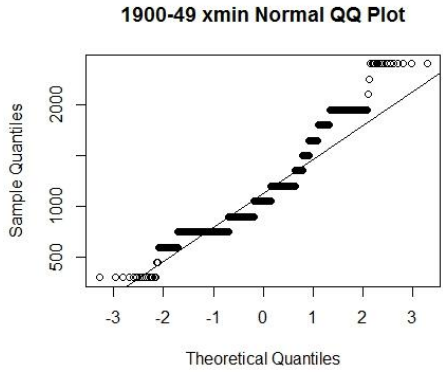
(b)



(c)

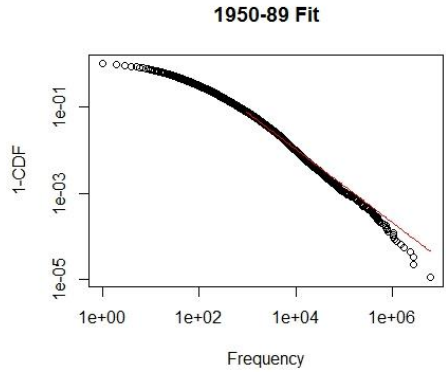


(d)

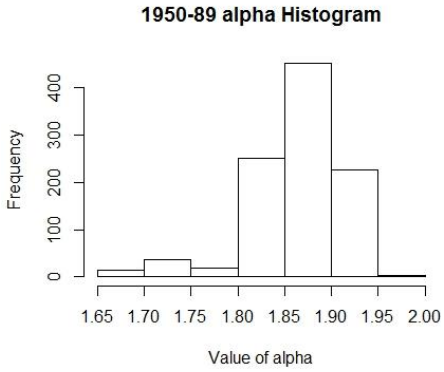


(e)

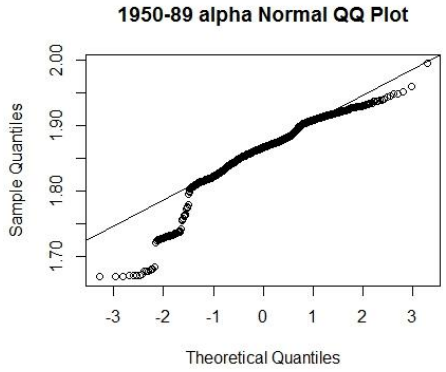
Figure 3: 1900-49 First Model. (a) Fit (b) alpha Histogram (c) alpha QQ Plot (d) xmin Histogram (e) xmin QQ Plot



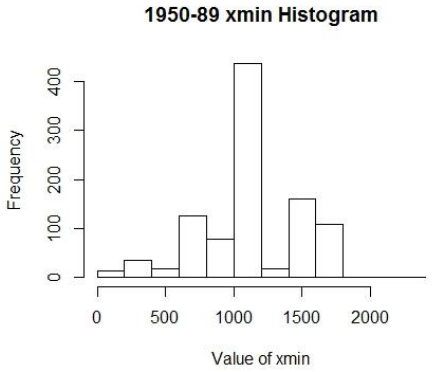
(a)



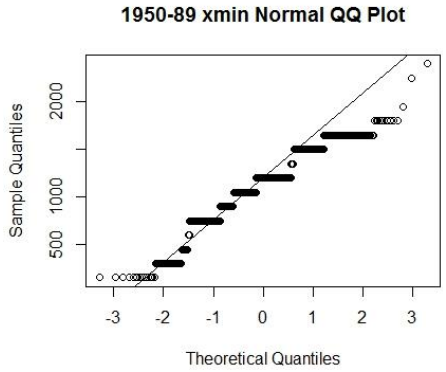
(b)



(c)

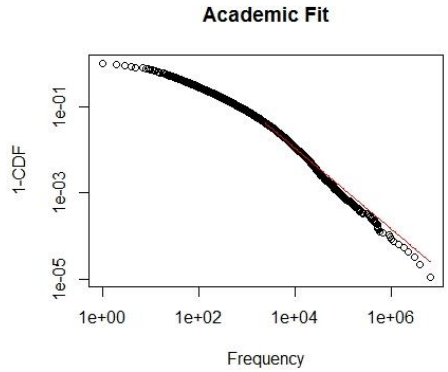


(d)

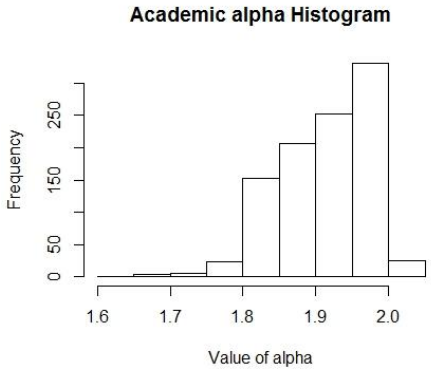


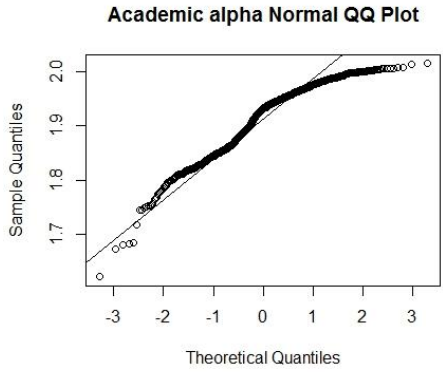
(e)

Figure 4: 1950-89 First Model. (a) Fit (b) alpha Histogram (c) alpha QQ Plot (d) xmin Histogram (e) xmin QQ Plot

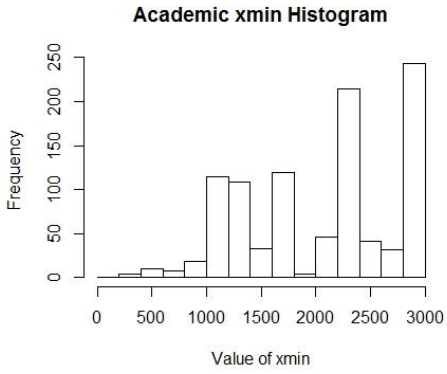


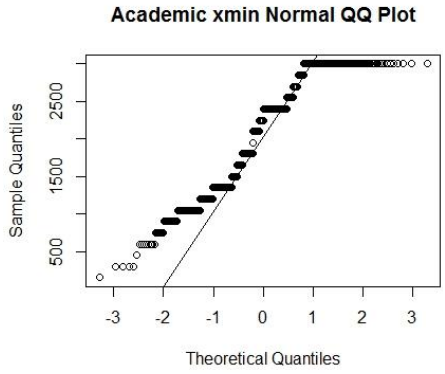
(a)





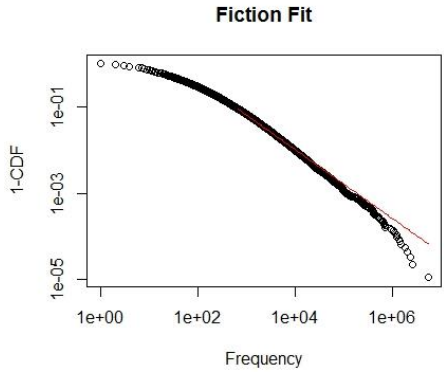
(b) (c)



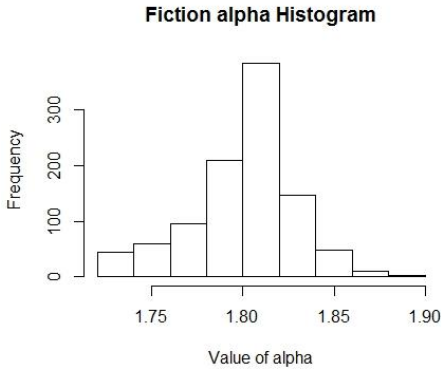


(d) (e)

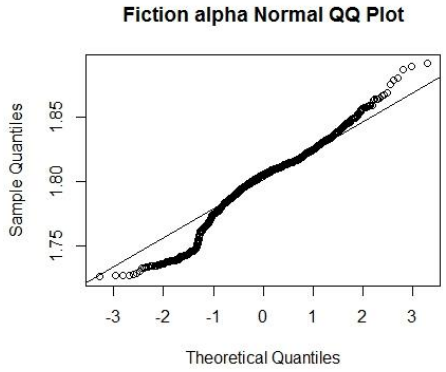
Figure 5: Academic First Model. (a) Fit (b) alpha Histogram (c) alpha QQ Plot (d) xmin Histogram (e) xmin QQ Plot



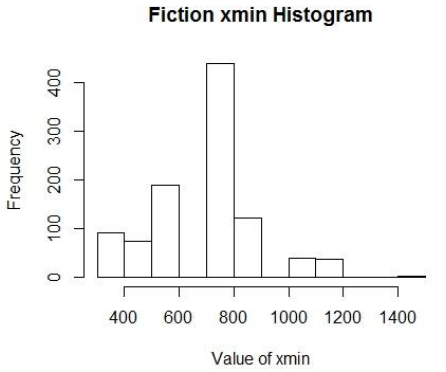
(a)



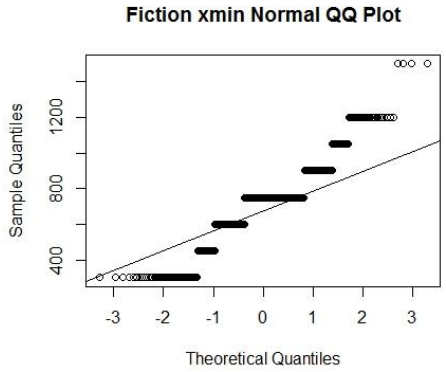
(b)



(c)

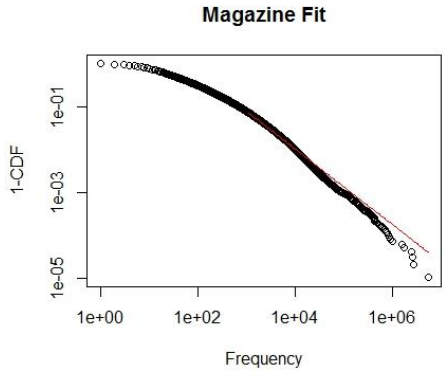


(d)

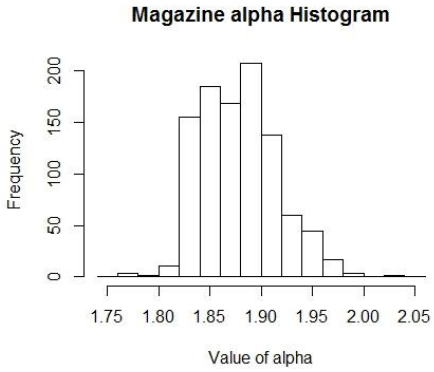


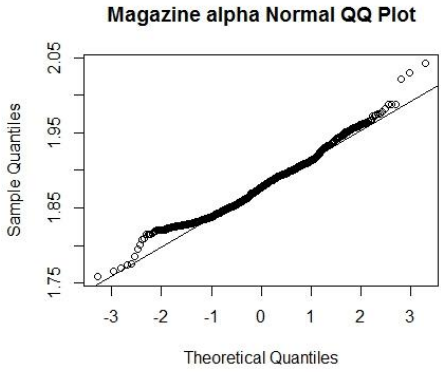
(e)

Figure 6: Fiction First Model. (a) Fit (b) alpha Histogram (c) alpha QQ Plot (d) xmin Histogram (e) xmin QQ Plot

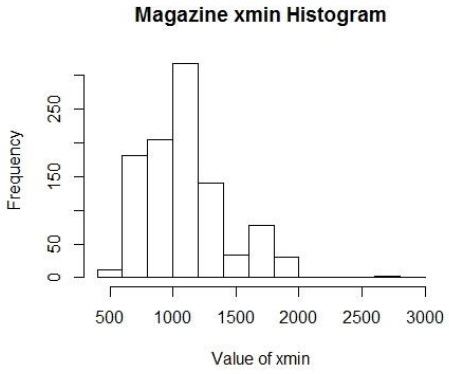


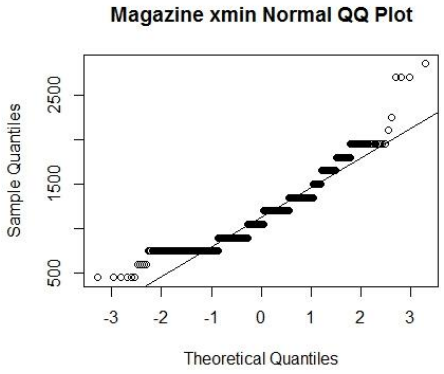
(a)





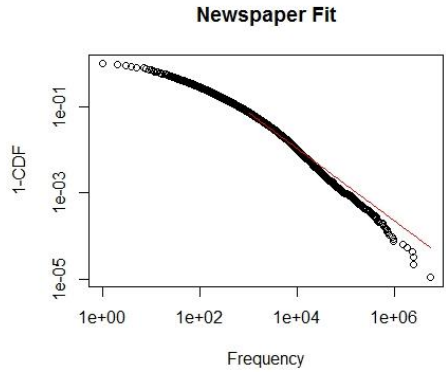
(b) (c)



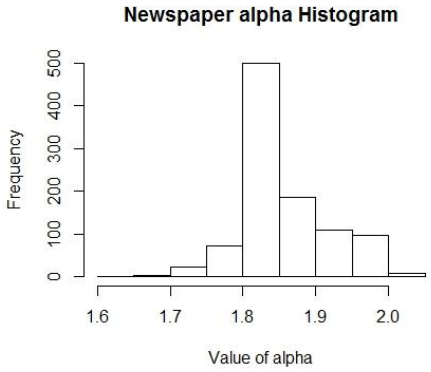


(d) (e)

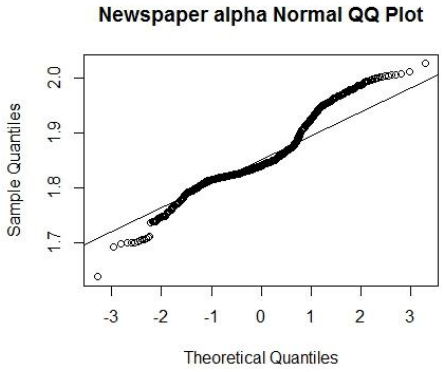
Figure 7: Magazine First Model. (a) Fit (b) alpha Histogram (c) alpha QQ Plot (d) xmin Histogram (e) xmin QQ Plot



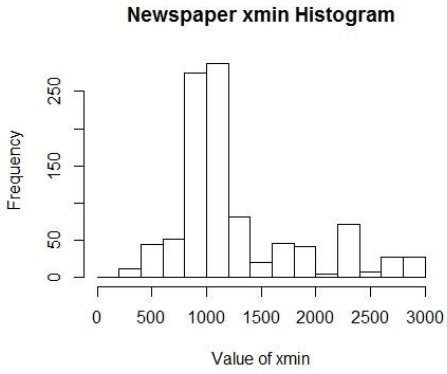
(a)



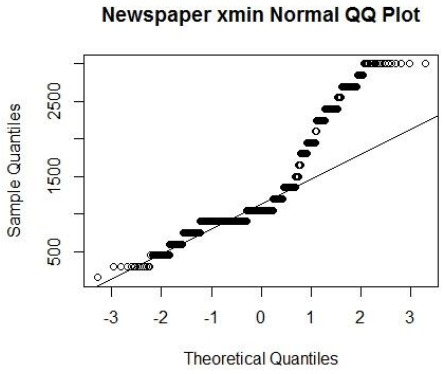
(b)



(c)

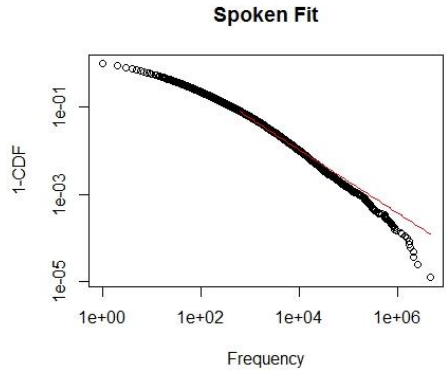


(d)

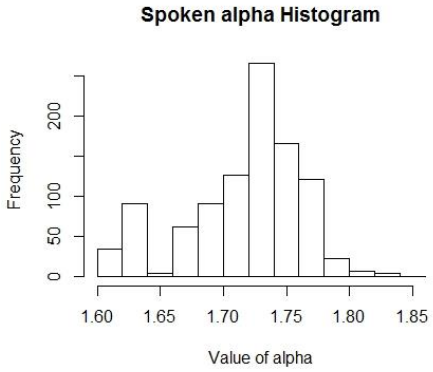


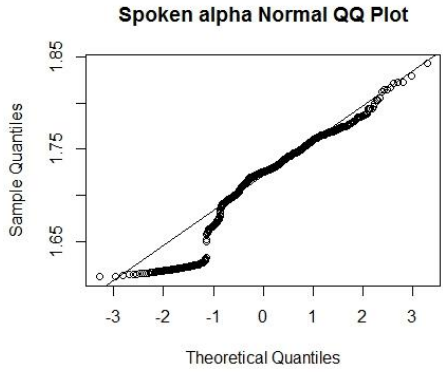
(e)

Figure 8: Newspaper First Model. (a) Fit (b) alpha Histogram (c) alpha QQ Plot (d) xmin Histogram (e) xmin QQ Plot

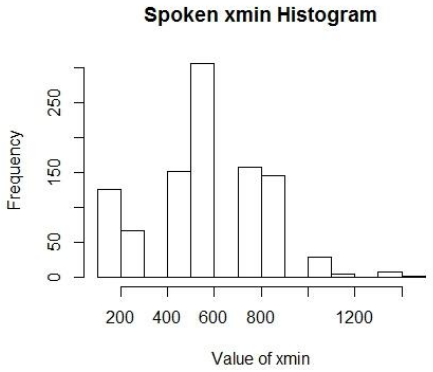


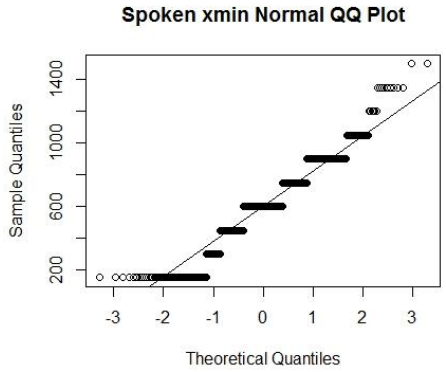
(a)





(b) (c)





(d) (e)

Figure 9: Spoken First Model. (a) Fit (b) alpha Histogram (c) alpha QQ Plot (d) xmin Histogram (e) xmin QQ Plot

4.3 First Model Analysis

These models are all quite poor. Beginning with the tables, problems immediately arise in Table 1, where one can see the low P-values (lower than 5%) for the Shapiro-Wilk tests. This means that the likelihood that the bootstrap iteration data for either the minimum rank cut-off value ("xmin") or the exponent of best fit ("alpha") are normal is very slim. This undermines any results from the F-tests and from the student-t tests since there's no basis for comparison between any of these parameters anymore.

Looking at the graphs, the results there are also disappointing. The histograms (parts (b) and (d) for Figures 2-9) for both xmin and alpha look nothing like bell-shaped curves, and the normal QQ plots (parts (c) and (e) for Figures 2-9) had points which deviated strongly from the "normal behavior" line. Perhaps most importantly, however, the fits (part (a) for Figures 2-9) don't look very good since the red line clearly sticks out and away from the actual locations of the data points.

The first model which was devised for this thesis was unacceptable. No useful inference could be gleaned since the models were nonsense. Then I noticed that, although the deviant behavior in the alpha parameters was bad (parts (b) and (d) for Figures 2-9), the deviations in the xmin values were far worse. Therefore, I decided on the following modification to the model, devising a second model that would yield more useful information.

In the tables of calculated minimum rank cut-off values, I noticed that the largest estimate was 2,395 for the "Academic" corpus (Table 1, column 2). Thus, any tokens that appeared less than 2,395 times were removed from consideration. This corresponds to any token of rank 4,037 or greater in a list that has 93,877 items in it (93,877 = 100,000

words minus all zero-frequency tokens) was removed. Thus, only the 4-5% most frequent of the words were actually being modelled. Looking at the graphs of all the model fits (part (a) in Figures 2-9), I saw that the points in all of the graphs were in a relatively straight line beyond a certain point. That is, the most frequent words tended to behave in a manner which most closely fit Zipf's Law – which is a straight line in a log-log graph.

Since only the high-frequency words seemed to have behavior which actually correlated to Zipf's Law, those were the only words which ought to have been modeled by Zipf's Law in the first place.

As justification for focusing only on these tokens to the exclusion of the rest, I noted that the most frequent words in a text tend to be function words (the, of, and, a, an, etc…) and not content words (elephant, ambidextrous, etc…) (Fries 1952). Variation in content words is something to be expected when observing different texts. Changes in the frequencies of high frequency words, however, would represent a more significant change in the deep structure of the language since function words (like pronouns, articles, and conjunctions) are elementary components of a language's grammar.

With these observations in mind, here is a short reiteration of the modifications which I made to create the second model: Firstly, instead of allowing the minimum rank cut-off value to float and be decided by the algorithm, I chose a value for the parameter for each corpus such that only the 3500 most frequent words in each corpus be modeled. In this way, the uncertainty in the minimum rank cut-off value is removed from the calculation and the exponent values are able to be compared with each other. Secondly, in addition to setting the minimum rank cut-off value for each corpus, the bootstrap

procedure was run for 5000 iterations instead of 1000 for this second set of models in order to better ensure that the bootstrap iterations are normally distributed.

4.4 Second Model Tables

The tables of the results of the tests and calculations for the second model are in this section. A description of the data presented in each table is given after each caption:

Corpus	Exponent of Best Fit	Standard Error 95%	Shapiro-Wilk Test P-Value
1800s	1.954	0.0156	0.159
1900-49	1.950	0.0157	0.200
1950-89	1.944	0.0158	0.773
Academic	1.962	0.0148	0.902
Fiction	1.873	0.0144	0.154
Magazine	1.984	0.0157	0.727
Newspaper	1.949	0.0148	0.159
Spoken	1.812	0.0128	0.627

Table 6: Exponent Estimates, Exponent Error Estimates, and Shapiro-Wilk P-Value

This table is analogous to Table 1 of the first model, in that it gives the point estimates of the parameters calculated by the algorithm, a 95% confidence interval, and a Shapiro-Wilk Test P-value. Unlike Table 1, there are only estimates for the exponent of best fit (“alpha”) since the minimum rank cut-off value (“xmin”) which is required by Clauset’s algorithm (2015) was chosen and set beforehand to be so that only the 3500 most frequent words would be modelled. Also unlike Table 1, here one can see that the P-values of the Shapiro-Wilk tests are all above 5%, so the hypothesis that the bootstrap iteration data (with 5000 iterations this time) is normal can be upheld. This is good because now there is a good reason to trust the test results on Tables 7 and 8.

Time-Period Corpus Pair	F Test Variance Ratio	Student T Test Mean Absolute Difference
1800s vs 1900-49	0.9890	0.0035
1800s vs 1950-89	0.9784	0.0092
1900-49 vs 1950-89	0.9893	0.0056
Media-Form Corpus Pair	**F Test Variance Ratio**	**Student T Test Mean Absolute Difference**
Academic vs Fiction	1.0598	0.0886
Academic vs Magazine	0.8867	0.0230
Academic vs Newspaper	0.9931	0.0120
Academic vs Spoken	1.3357	0.1496
Fiction vs Magazine	0.8367	0.1116
Fiction vs Newspaper	0.9371	0.0766
Fiction vs Spoken	1.2604	0.0610
Magazine vs Newspaper	1.1201	0.0350
Magazine vs Spoken	1.5064	0.1725
Newspaper vs Spoken	1.3449	0.1376

Table 7: F Test Ratios, Student T Test Differences for Exponent of Best Fit

Table 7 is much like Table 4 and is read in exactly the same way. The primary difference is that here one can see that the variance ratios are overall much closer to equaling 1. In addition, there is a clear distinction between the time-period data and the media-form data which wasn't apparent in the first model: The variance ratios for the time-period data are much more uniformly close to 1 than the variance ratios for the media-form data. In addition, the mean absolute differences in the third column show a similar distinction, in that the differences in the time-period corpora tend to be at least ten times smaller than the differences in the media-form data. That is, there are two zeros after the decimal for the time-period corpus pair mean absolute differences, whereas there's only at most one zero after the decimal for the media-form corpus pairs. This, alone, is enough to make a conclusion for the thesis: The parameters for media-form data differ much more strongly from each other than the parameters for the time-period data.

Time-Period Corpus Pair	F Test P-Value	Student T Test P-Value
1800s vs 1900-49	0.6966	<2.2e-16
1800s vs 1950-89	0.4403	<2.2e-16
1900-49 vs 1950-89	0.7027	<2.2e-16
Media-Form Corpus Pair	**F Test P-Value**	**Student T Test P-Value**
Academic vs Fiction	0.0402	<2.2e-16
Academic vs Magazine	2.129e-05	<2.2e-16
Academic vs Newspaper	0.8076	<2.2e-16
Academic vs Spoken	<2.2e-16	<2.2e-16
Fiction vs Magazine	2.985e-10	<2.2e-16
Fiction vs Newspaper	0.0219	<2.2e-16
Fiction vs Spoken	4.441e-16	<2.2e-16
Magazine vs Newspaper	6.131e-05	<2.2e-16
Magazine vs Spoken	<2.2e-16	<2.2e-16
Newspaper vs Spoken	<2.2e-16	<2.2e-16

Table 8: F Test P-Values, Student T Test P-Values for Exponent of Best Fit

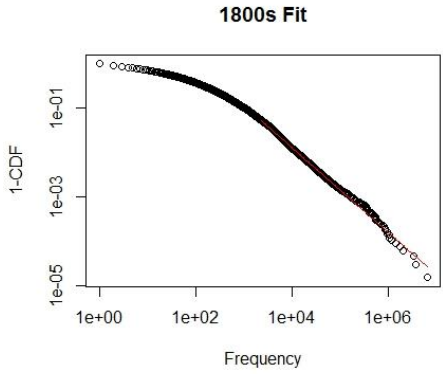
Table 8 is much like Table 5 from the first model, but there is one important difference. Here, the meaning of the student-t test is different. All of the P-values in the far-right column are very low (essentially, too low to make any sort of comparison), indicating that the hypothesis that the means of the paired corpora's bootstrap iterations are equal is quickly rejected. The reason why that's not a concern is that the 5000 iterations for each corpus alpha parameter produces so much evidence that each individual corpus alpha parameter is equal to a specific value. In other words, to compare them and ask the question, "Are these equal or not?" results in the test immediately responding with "No, they aren't equal." Essentially, there's ample evidence that the alpha for the 1800s corpus is equal to X and there's ample evidence that the alpha for the Academic corpus is equal to Y. So the student t-test P-value is not so useful.

What are useful, however, are the P-values of the F-tests. One can see that the P-values for the F-tests of the time-period corpora pairs are clearly higher than 5%, and that's good, but the P-values for the F-tests of the media-form corpora pairs clearly vary a

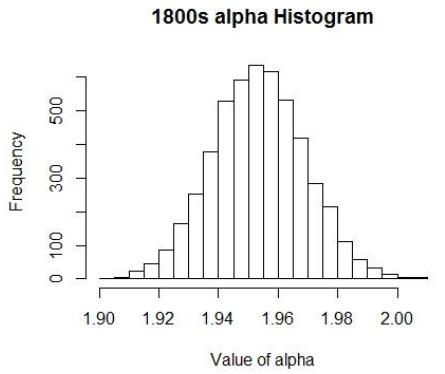
lot. Furthermore, only one pair – the Academic vs. Newspaper pair – has a P-value that could be considered high enough. This lends more credence to the idea that the media-form data vary more between themselves than the time-period data do.

4.5 Second Model Graphs

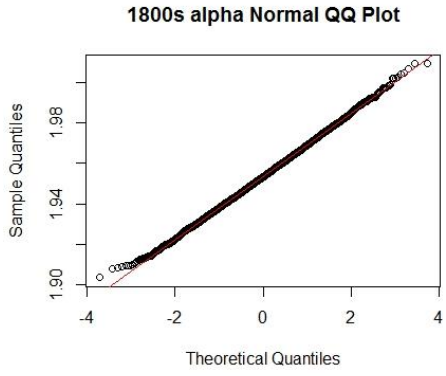
The graphs of the second round of models (Figures 10-17) are organized so that there are graphs for two corpora to a page (there are two Figures for each page), with one corpus taking up the left hand side and the other corpus taking up the right hand side. Each Figure has three graphs: part (a), part (b), and part (c). The top graphs (part (a)) are the Zipf's Law fits as decided by Clauset's algorithm (2015). They are read in the same way as they were for the first model, with the black circles representing individual data points and dark black indicating a dense cluster of data points. The red line is the fitted model. Part (b) of each figure is the histogram of bootstrap iterations for the alpha parameter of each corpus. In the same way as before, the target shape is a "bell-curve" because normality of this data is desirable, and normality is desirable because that is how there is a basis for comparison between the calculated alpha parameters of each corpus. Part (c) is the normal QQ plot of the bootstrap iteration data, and as before, the desired outcome is for the data points to be as close as possible to the normal-behavior line.



(a)

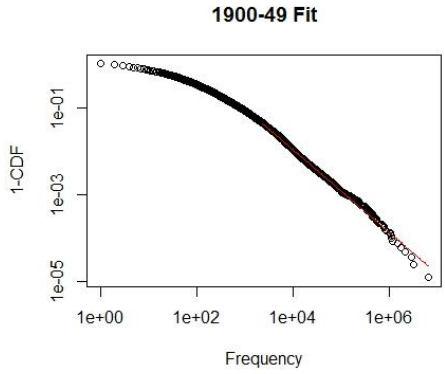


(b)

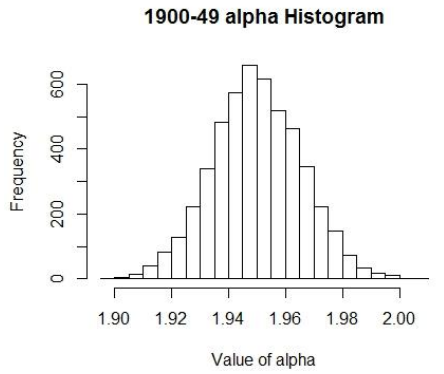


(c)

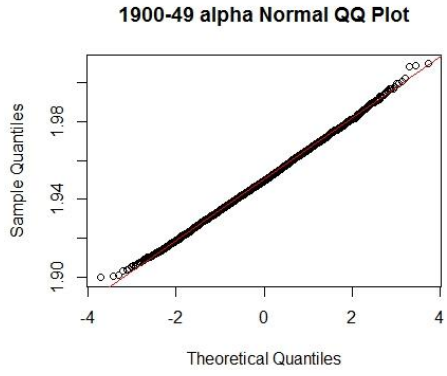
Figure 10: 1800s Second Model (a) Fit (b) alpha Histogram (c) alpha QQ Plot



(a)

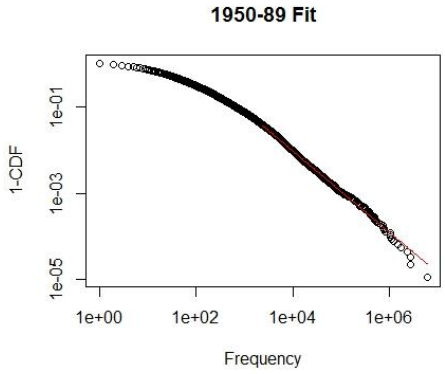


(b)

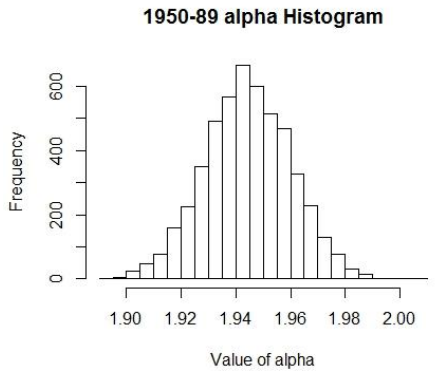


(c)

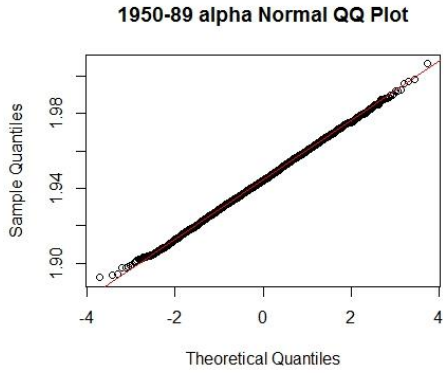
Figure 11: 1900-49 Second Model (a) Fit (b) alpha Histogram (c) alpha QQ Plot



(a)

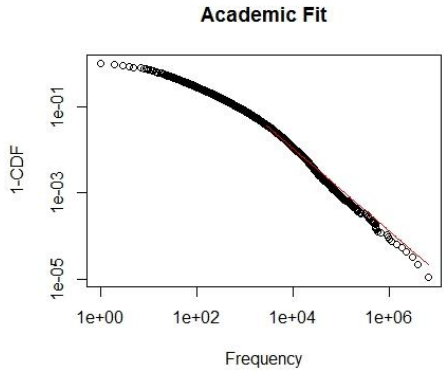


(b)

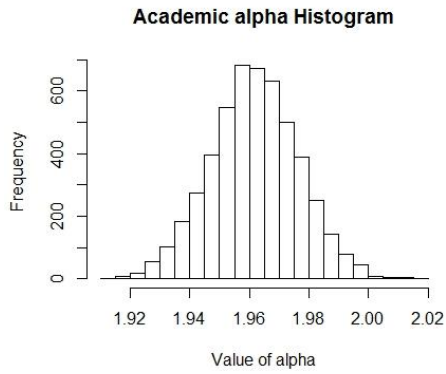


(c)

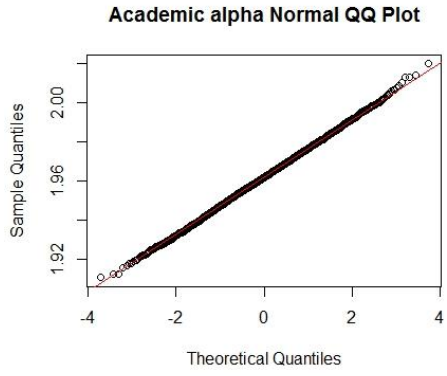
Figure 12: 1950-89 Second Model (a) Fit (b) alpha Histogram (c) alpha QQ Plot



(a)

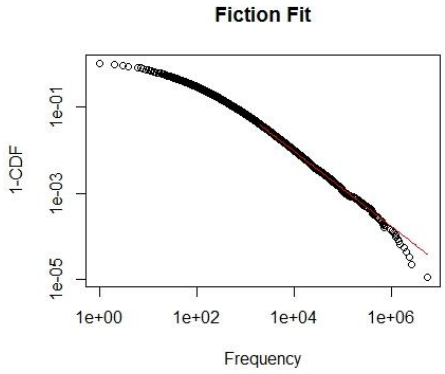


(b)

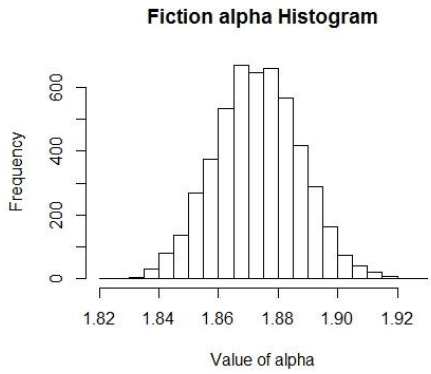


(c)

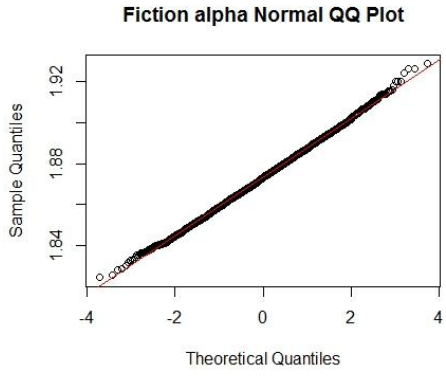
Figure 13: Academic Second Model (a) Fit (b) alpha Histogram (c) alpha QQ Plot



(a)

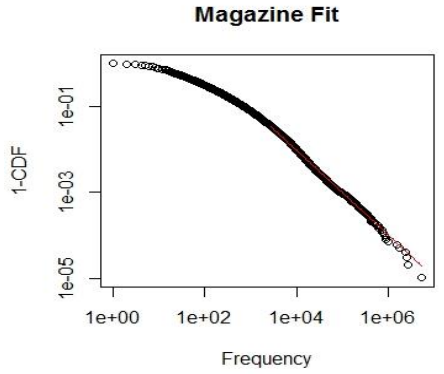


(b)

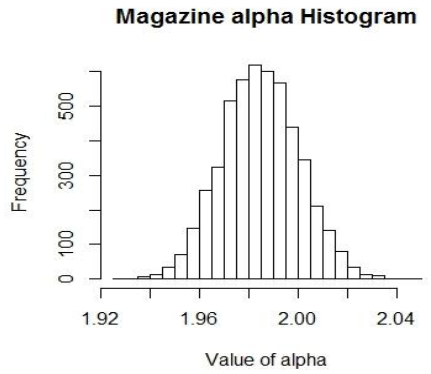


(c)

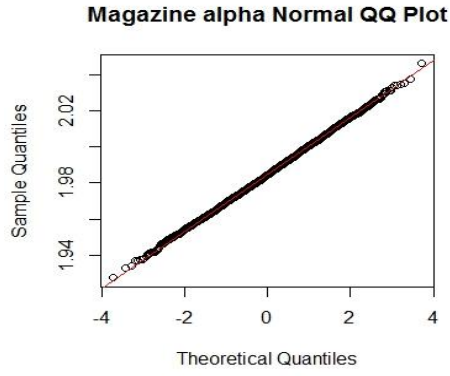
Figure 14: Fiction Second Model (a) Fit (b) alpha Histogram (c) alpha QQ Plot



(a)

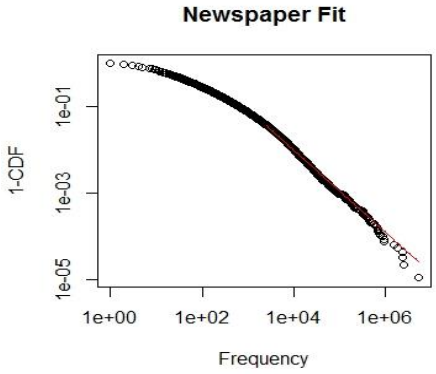


(b)

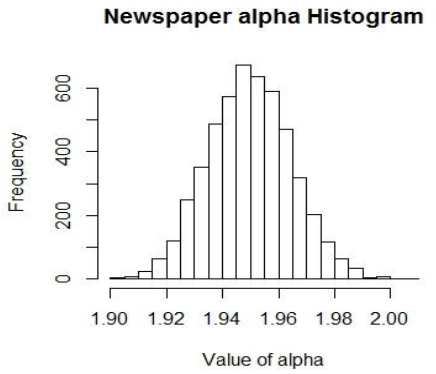


(c)

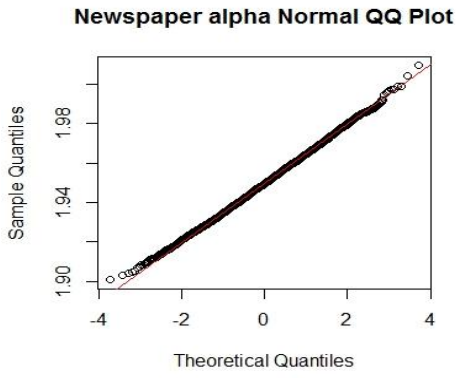
Figure 15: Magazine Second Model (a) Fit (b) alpha Histogram (c) alpha QQ Plot



(a)

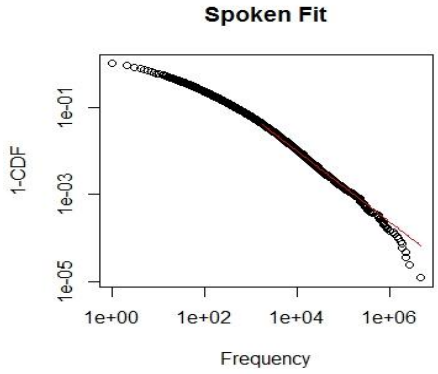


(b)

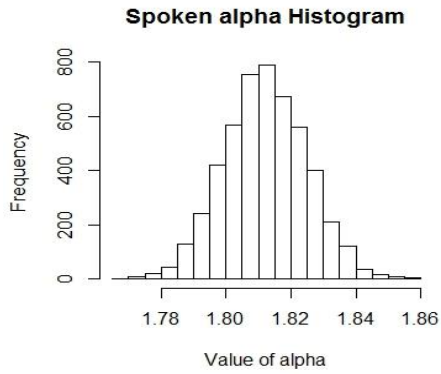


(c)

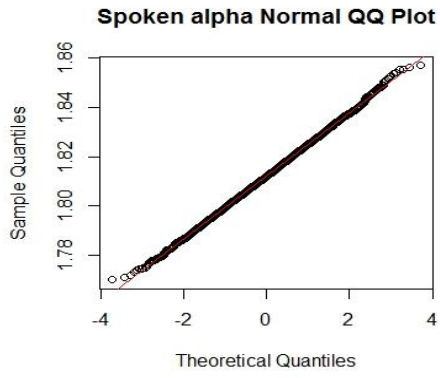
Figure 16: Newspaper Second Model (a) Fit (b) alpha Histogram (c) alpha QQ Plot



(a)



(b)



(c)

Figure 17: Spoken Second Model (a) Fit (b) alpha Histogram (c) alpha QQ Plot

4.6 Second Model Analysis

On account of the much better results from the Shapiro-Wilk tests on Table 6, the clear bell-shaped curves in the histograms (part (b) of Figures 10-17), and the close proximity of the points to the normal-behavior line in the normal QQ plots (part (c) of Figures 10-17), the second model is much more trustworthy than the first model. In addition, by looking at the fits (part (a) of Figures 10-17), one can see that the model is much closer to the data than before. Because of this and the observations made in the comments on the Tables 6-8, the conclusion of the thesis can now be made.

5 Conclusions

In conclusion, the second round of tests shows that the Zipf's Law fits for corpora of texts which are selected from different media forms tend to vary more than the fits for the corpora selected according to time period. The F Test Variance Ratios in Table 7 are the key to concluding in favor of this alternative hypothesis. Here, one can see that the variances in the bootstrap iterations between the time-period corpora are reasonably close to 1. However, by comparison, the media-form corpora pairs have variance ratios which differ greatly. This effect can also be seen the Absolute Mean Difference column of Table 7, where it can be seen that the means of the bootstrap iterations for two different corpora of media-form data tend to have a difference which is at least ten times greater than a difference between the means of the bootstrap iterations for two time-period corpora.

There are several possible directions for future research from here, and three principal directions will be described. First, it's clear from looking at any of the log-log graphs of the data that Zipf's Law – which, on a log-log graph, is a straight line – is only

particularly good at modelling the behavior of the most frequent words of these corpora. Since it was necessary to fix the minimum-rank cut-off to a common value for each corpora in order to get useful inference, there's a clear need to model the remainder of the non-high frequency data using something other than Zipf's Law. Therefore, a future direction is to determine a model for the remainder of the data on the left-hand side of the graphs representing less frequent words. Putting such a model together with the Zipf's Law fit in a piecewise graph would form a more accurate, complete model for the data.

Secondly, the manner in which the COCA and the COHA data was collected should be further investigated. Being able to source which texts the frequency data was taken from is important to the integrity of the model. For example, in the time period data for the 1800s, suppose that there is a word which incidentally is only found in newspapers or non-fiction books from that era, but not magazines or fiction novels. Then there is a need to account for how the word frequency is affected by the media-form variable or not.

Lastly, the goodness of the model fit to the data could stand to be more closely examined. There was no examination of the P-value of the Kolmogorov-Smirnoff statistic (Chakravarti 1967) itself for either the first model or the second model. This would be relevant information to see, because it would offer inference on how well the model fits the data. In addition, simply taking the difference between each point and the model, squaring those differences, and then adding those squares together would be a simple way to see and compare the quality of the model fits for each corpus to each other.

6 References

Baayen, R. (2001). *Word Frequency Distributions* (Vol. 1). Berlin: Springer

Calude, A. S., & Pagel, M. (2011). *How do we Use Language? Shared Patterns in the Frequency of Word Use across 17 World Languages.* Philosophical Transactions of the Royal Society B: Biological Sciences, 366(1567), 1101–1107.

Chakravarti, I., Laha, R., & Roy, J. (1967). Handbook of Methods of Applied Statistics, volume 1. John Wiley and Sons.

Clauset, A., Shalizi, C. R., & Newman, M. E. J. (2009) *Power-law distributions in empirical data*. SIAM Review, 51(4): 661–703.

Davies, M. (2008). *The Corpus of Contemporary American English: 450 million words, 1990-present*. Available online at http://corpus.byu.edu/coca/.

Davies, M. (2010). *The Corpus of Historical American English: 400 million words, 1810-2009.* Available online at http://corpus.byu.edu/coha/.

Fries, C. C. (1952). *The Structure of English*. New York: Harcourt Brace.

Gelbukh, A. & Sidorov, G. (2001). Zipf and Heaps Laws' Coefficients Depend on Language. In D. Hutchison (eds.), *Lecture Notes in Computer Science* (vol. 2004) (332-335). Berlin: Springer.

Gillespie, C. S. (2015). *Fitting Heavy Tailed Distributions: The poweRlaw Package.* Journal of Statistical Software, 64(2), 1-16.
URL http://www.jstatsoft.org/v64/i02/.

Larsen, R. J. & Marx, M. L. (2006). *An Introduction to Mathematical Statistics and Its Applications* (4th ed.). Upper Saddle River: Pearson Prentice Hall.

Lehmann, E. L. (1993). *The Fisher, Neyman–Pearson Theories of Testing Hypotheses: One Theory or Two?* Journal of the American Statistical Association 88(424): 1242–1249

Lu, L., Zhang, Z. & Zhou, T. (2013). Deviation of Zipf's and Heaps' Laws in Human Languages with Limited Dictionary Sizes. Sci. Rep. 3, 1082; DOI: 10.1038/srep01082.

Marcus, M. P., Marcinkiewicz, M. A., & Santorini, B. (1993). *Building a Large Annotated Corpus of English: The Penn Treebank*. Computational linguistics, 19(2), 313–330.

Piantadosi, S.T. (2014). *Zipf's Word Frequency Law in Natural Language: A Critical Review and Future Directions*. Psychometric Bulletin and Review, 21(5), 1112–1130.

R Core Team. (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/

Zipf, G. (1936). *The Psychobiology of Language*. London: Routledge.

Zipf, G. (1949). *Human Behavior and the Principle of Least Effort*. New York: Addison-Wesley.